Two types of Twitter users with equally many followers

Kodai Saito and **Naoki Masuda**
Department of Mathematical Informatics,
The University of Tokyo,
7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan

Abstract

The number of followers is acknowledged as the presumably most basic popularity measure of Twitter users. However, because it is subjected to manipulations and therefore may be deceptive, some alternative methods for ranking Twitter users that take into account users' activities such as the tweet and retweet rate have been proposed. In the present work, we take a purely network approach to this fundamental question. First of all, we show that there are two types of users possessing a large number of followers. The first type of user follows a small number of others. The second type of user follows almost as equally many others as the number of its followers. Such a distinction is prominent for Japanese, Russian, and Korean users among the seven language groups that we examined. Then, we compare local (i.e., egocentric) followership networks around the two types of users with many followers. We show that the latter type, which is presumably uninfluential users despite its large number of followers, is characterized by high link reciprocity, large clustering coefficient, a large fraction of the second type of users among the followers, and a small PageRank. We conclude that the number of others that a user follows is as equally important as the number of followers when estimating the importance of a user in the Twitter blogosphere.

Introduction

A prominent feature of social microblogging services including Twitter is that users can follow or subscribe specific other users whose activities are of interest. The number of followers is conventionally used as a succinct popularity measure of Twitter users (Ghosh et al. 2012). This quantity is shown on the profile webpage of each user, which makes it even popular. In addition, main activity-related measures of users such as the retweet rate are known to be also proportional to the number of followers of a user (Suh et al. 2010).

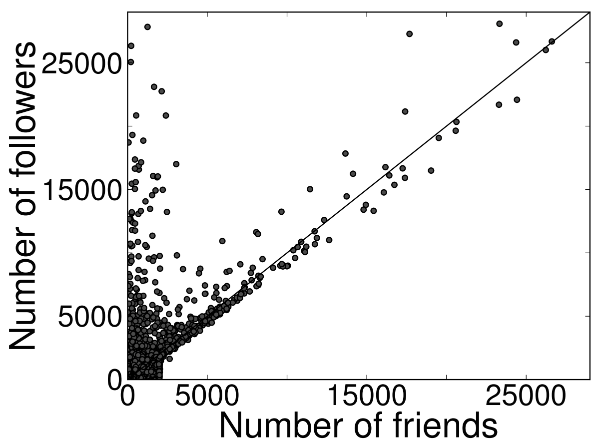
In the present paper, we propose that Twitter users with many followers are really popular only when they follow a small number of other users. The present study is motivated by the observation that there are two distinct types of users with many followers in some countries. The number of followers is plotted against the number of friends (i.e., those whom a user follows) for randomly sampled Japanese users in Figure 1(a). First, the figure indicates that only a small fraction of users have a large number of followers or friends, which is the stylized scale-free property present in various networks including Twitter's social networks (Ghosh et al. 2012; Weng et al. 2010; Kwak et al. 2010). Second, among users possessing many followers, some users follow a small number of others, whereas other users follow many others. In fact, the latter type of users has almost the equally large numbers of followers and friends. The number of friends cannot be much larger than the number of followers because of the restriction imposed by Twitter. Therefore, it is obvious that users are absent far off below the diagonal in Figure 1(a). Nevertheless, it is surprising that many users are concentrated near the diagonal and we find few users with many followers and intermediate numbers of friends.

Figure 1(b) is the density plot that magnifies Figure 1(a). We use the density plot because there are many users with small numbers of followers and friends. In this region, there is no system restriction on the number of followers and that of friends; any user is allowed to possess up to 2000 followers and friends. Figure 1(b) indicates that many users are concentrated on the diagonal, which is consistent with the results shown in Figure 1(a).

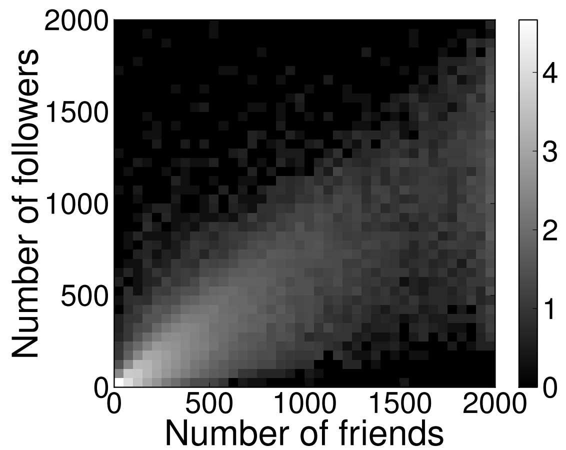
Figure 1 indicates that the number of followers may not be a good popularity measure of users. The same claim has been made on the basis that the number of followers is easily manipulated by link farming and spammer activities and the following may not directly reflect activities of the followers. Therefore, alternative popularity measures may be more useful (Weng et al. 2010; Cha et al. 2010; Bakshy et al. 2011).

The correlation between the number of followers and that of friends is shown in a previous study (Java et al. 2007), but not as strong as that implied in Figure 1. In 2007, Twitter was much less known than it is now. Therefore, their data and contemporary data including ours can be different in demography. In particular, Twitter is now used in various countries, and its usage may depend on countries. Therefore, we decided to sample local (i.e., egocentric) networks of Twitter users separately for some major countries, where the classification is based on the language and location of

the users. We quantify differences between local followership networks around two types of users using five quantities and the PageRank. Based on the results we argue that, although the two types of users have similar numbers of followers, they receive the follow of different qualities. In other words, those having an equally large number of followers and friends may not be really important even if they enjoy a huge number of followers.



(a)



(b)

Figure 1: (a) Number of friends and that of followers for sampled Japanese users of Twitter. A dot represents a user. The diagonal line is shown as a guide to the eyes. (b) Density plot of the number of friends and that of followers for Japanese users with less than 2000 friends and followers.

Data sets

Twitter is a major microblogging service that started to operate on July 2006 and enjoys more than 5×10^8 registered users as of July 2012. Users of Twitter can send and read text message of up to 140 characters called "tweet". Users can read tweets of other users by registering their accounts, i.e., by following them. The population of users constitutes a directed network in which a link emanates from the follower to the followee.

We mainly analyze local networks around specified users registering either of the seven languages, i.e., English, Spanish, Japanese, Portuguese, Russian, Korean, and French. We selected the seven languages because each language is used by a sufficient number of users such that language-wise statistical analysis is possible. In general, users are connected with those registering the same language with a larger probability than with those registering different languages (Takhteyev, Gruzd, and Wellman 2012). Therefore, the local network of a selected user tends to be homogeneous in terms of the language.

By using Twitter representational state transfer application programming interface (API) (Russell 2011), we acquired properties of users including the number of followers (**followers_count**), the number of friends (**friends_count**, i.e., the number of users that a user follows), and the language (**lang**). The operating institution allows general users including us to collect the Twitter users' network at a limited speed. We registered an application of Twitter as a developer and authenticated the application by the OAuth 2.0 protocol to use the so-called **users/lookup**, **followers/ids**, and **friends/ids** resources. The **followers/ids** and **friends/ids** resources return error when the targeted users protect their tweets and are not followed by our test account. To acquire IDs of friends and followers of such protected users, we would have to beg them to accept our following. Therefore, we excluded the protected users, which account for 1–10% of the entire users, from the following analysis.

We are concerned with local networks of users with relatively many followers. We sample such users by the two resources called the neighbor sampling and random sampling methods defined as follows.

In the neighbor sampling, we first select seed users and then sample followers of the seed users. It should be noted that we are not interested in the seed users. We define users with many followers, as identified by the "twitaholic" website (*http://twitaholic.com/*), as seed users, to realize a large sample size. To this end, for seven countries where the corresponding languages are spoken as the dominant official language (i.e., US, Spain, Japan, Brazil, Russia, Korea, and France), we collect users whose residence location property contains the name of the city with the largest population in the country. Then, for each country, we select three users as seeds such that they are not accounts created by an organization or company and they have the largest number of followers among those having less than 5×10^5 followers. We exclude users with more than 5×10^5 followers because we will exhaustively collect the IDs of their followers, and the API does not allow us to collect users' data at a sufficiently high speed. Then, we acquire the IDs of the seeds' all followers. The restriction of the API makes it difficult to collect the local networks of all the seeds' followers. Therefore, we randomly select 5×10^4 users among the seeds' followers and

acquire their properties when the following analysis requires local networks of users. Finally, homophily with respect to the language implies that the seeds' followers tend to register the same language as that of the seed user. Because we will separately analyze users for different language groups, we excluded users registering a different language from that used by the seed user.

In the random sampling, we randomly create 1.5×10^6 IDs as uniformly and independently distributed integers between 12 (corresponding to the first user) and the maximum ID value among those of the seeds' followers identified by the neighbor sampling. Because seeds are often popular and followed by new users, the uniform distribution defined in this way approximates the unbiased random sampling of a user. Finally, we sift out the users registering either of the seven target languages.

We use the two sampling methods for the following reasons. First, with the neighbor sampling, a sampled user has a much larger number of followers on average because the users are sampled conditioned that they follow somebody. Therefore, the neighbor sampling allows us to investigate the statistics of users having many followers as compared to the random sampling does. Second, with the neighbor sampling, properties of the sampled users may be correlated because a large fraction of them follows the same seed user. The random sampling method does not suffer from such correlation.

We do not filter users according to their activities except that the IDs banned by Twitter or deleted by users are neglected. Our samples may contain spammers. Nevertheless, at least the users collected by the neighbor sampling are mostly not spammers because they follow a celebrity user by definition. Up to our manual inspections, most users collected by either sampling method are not spammers.

The sample sizes for the different sampling methods and languages are summarized in Table 1. All the data are retrieved between October 12, 2012 and January 11, 2013.

Table 1: Number of users sampled by the neighbor and random sampling methods.

Language	Neighbor sampling	Random sampling
English	118316	638122
Spanish	129415	126350
Japanese	113140	44204
Portuguese	95211	43353
Russian	70354	24940
Korean	48367	13636
French	51571	22821

Results

Users having approximately many followers and friends

In Figure 1, we showed that some users have similar k^{in} (i.e., number of followers) and k^{out} (i.e., number of friends) values. To scrutinize this observation, we measure two indices for each language group. First, we define the degree ratio by

$$r = \left\langle \frac{\min(k^{\mathrm{in}}, k^{\mathrm{out}})}{\max(k^{\mathrm{in}}, k^{\mathrm{out}})} \right\rangle \quad (1)$$

where $\langle \cdot \rangle$ represents the average over the users in a language group. If k^{in} and k^{out} are close for many users, r is large. Second, we define the diagonal fraction, denoted by d, as the fraction of users that satisfy

$$k^{\mathrm{out}}/1.1 \leq k^{\mathrm{in}} \leq 1.1 \times k^{\mathrm{out}}. \quad (2)$$

Both r and d range between 0 and 1. The r and d values may be strongly affected by users having small k^{in} and k^{out} values, which occupy the majority owing to the long-tailed distributions of k^{in} and k^{out} (Ghosh et al. 2012; Weng et al. 2010; Kwak et al. 2010). Because in this study we focus on properties of users having relatively many friends and followers, we restrict ourselves to the users satisfying $k^{\mathrm{in}}, k^{\mathrm{out}} > 100$ or $k^{\mathrm{in}}, k^{\mathrm{out}} > 2000$.

The r and d values for the different sampling methods, language groups, and threshold degrees (i.e., 100 or 2000) are shown in Table 2. Regardless of the sampling method and threshold degree, r and d are large for the Japanese, Russian and Korean groups, intermediate for the English group, and small for the Spanish, Portuguese, and French groups. Therefore, the observation that many users have similar indegree and outdegree, as shown in Figure 1 for Japanese users, is eminent for Japanese, Russian and Korean among the seven languages.

Table 2: Degree ratio (r) and the diagonal fraction (d) for the users satisfying $k^{\mathrm{in}}, k^{\mathrm{out}} > 100$ (values left to the slash) and 2000 (values right to the slash).

Language	r(neighbor)	r(random)	d(neighbor)	d(random)
English	0.299/0.532	0.429/0.415	0.031/0.209	0.080/0.180
Spanish	0.360/0.257	0.395/0.399	0.031/0.050	0.059/0.179
Japanese	0.585/0.635	0.695/0.722	0.115/0.333	0.250/0.473
Portuguese	0.232/0.315	0.386/0.342	0.013/0.049	0.051/0.090
Russian	0.408/0.759	0.409/0.627	0.091/0.517	0.074/0.500
Korean	0.439/0.752	0.598/0.824	0.072/0.548	0.218/0.685
French	0.313/0.464	0.379/0.238	0.028/0.169	0.048/0.036

Two types of users with many followers

Definition of type 1 and 2 user Our main hypothesis is that the quality of the follow may be different between users with large k^{out} and those with small k^{out} even if the users enjoy equally many followers (i.e., large k^{in}). To investigate this issue on the basis of the followership network, we classify users with many followers into two types as follows (Figure 2). We define type 1 users as those satisfying $2500 \leq k^{\mathrm{in}} \leq 7500$ and $k^{\mathrm{out}} \leq 500$. Type 1 users are followed by many users and do not follow many others. We define type 2 users as those satisfying $k^{\mathrm{out}}/1.1 \leq k^{\mathrm{in}} \leq 1.1 \times k^{\mathrm{out}}$ and $5000 \leq k^{\mathrm{in}} + k^{\mathrm{out}} \leq 15000$. Type 2 users are followed by many users and follow many others. The operating institution of Twitter does not allow users with $k^{\mathrm{out}} \geq 2000$ to

have more than $k^{\mathrm{out}} \geq 1.1 \times k^{\mathrm{in}}$ friends. Therefore, many users are located near the diagonal in Figure 1 partly owing to the system restriction. Nevertheless, we are interested in the behavior of type 2 users.

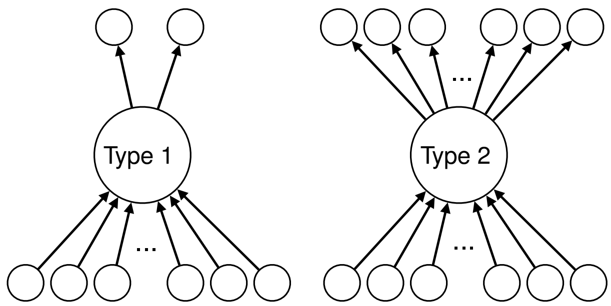


Figure 2: Schematic of the two types of users with many followers

Indegree k^{in} of type 2 users is distributed on roughly the same range as k^{in} of type 1 users (i.e., $2500 \leq k^{\mathrm{in}} \leq 7500$). Therefore, type 1 and 2 users are statistically indifferent in terms of k^{in}. We may be able to reveal the difference between the two types of users by inspecting contents of the tweets and other activities of these users (e.g., tweet and retweet rates). In the following, we take a complementary, purely network-based approach. In the remainder of this section, we compare type 1 and type 2 users by examining five quantities derived from their local networks.

Local link reciprocity of type 1 and 2 users First, we hypothesize that type 2 users have many followers because the type 2 users follow back their followers to keep them around (i.e., reciprocal links). To prove this, we define local link reciprocity (reciprocity for short), of a type 1 or 2 target user as the number of the target user's friends that follow back the target user, normalized by k^{out} of the target user. Because k^{out} is dissimilar between type 1 and 2 users by definition, the reverse definition, i.e., the number of the target user's followers that a target user follows back, normalized by k^{in} of the target user, is invalid. The upper bound of the latter quantity is much smaller for type 1 users than type 2 users.

For each language group, the mean and standard deviation of the reciprocity of the ten randomly selected users of type 1 or 2 are shown in Table 3. The table indicates that type 2 users have significantly larger reciprocity than type 1 users, at least for the Japanese, Russian, and Korean groups, for which the distinction between the type 1 and 2 users is clear (Table 2). It should be noted that approximately 80 % of links in Twitter are reciprocal (Weng et al. 2010) (but see (Cha et al. 2010)). This is consistent with the results shown in Table 3, in which the reciprocity values are generally large.

Outdegree of those following a type 1 or 2 user Second, we examine k^{out} (i.e., number of friends) for those following a type 1 or 2 user (Figure 3(a)). If k^{out} is large, the follow that a type 1 or 2 user receives may not be valuable because the amount of time that a follower spends on looking at others' tweets would be inversely proportional to k^{out} to the first-order approximation.

Table 3: Local link reciprocity for different language groups.

Language	Type 1	Type 2
English	0.364±0.240	0.656±0.230
Spanish	0.478±0.181	0.669±0.192
Japanese	0.600±0.206	0.872±0.102
Portuguese	0.280±0.234	0.420±0.233
Russian	0.452±0.185	0.861±0.232
Korean	0.648±0.214	0.884±0.069
French	0.557±0.235	0.851±0.196

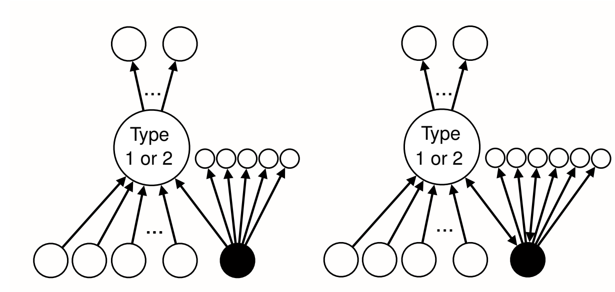
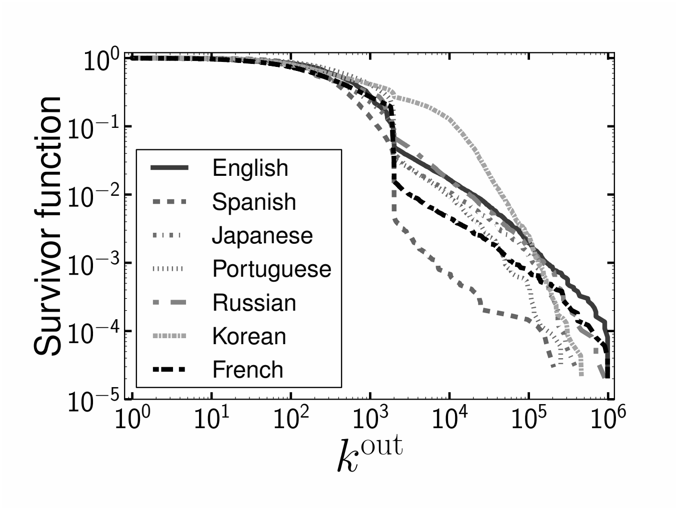


Figure 3: (a) Outdegree of those following a type 1 or 2 user. It is equal to 6 for the user shown by the filled circle. (b) Follower's reciprocity. It is equal to $2/7$ for the user shown by the filled circle.

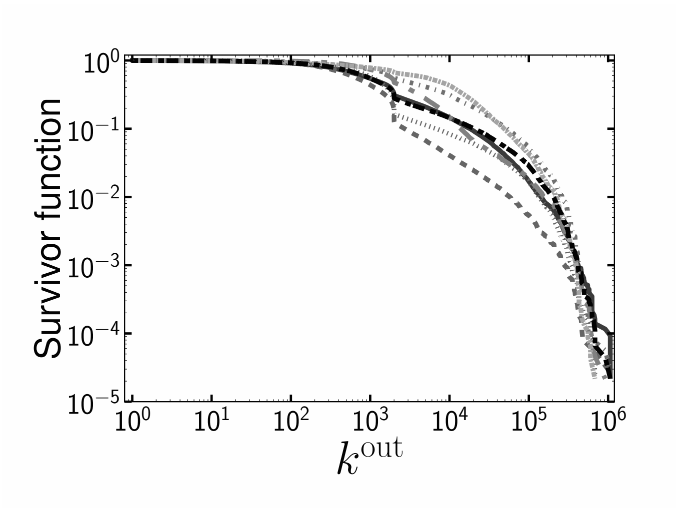
For those that follow any of the ten selected type 1 or 2 users of each language, the survivor functions of k^{out} (i.e., fraction of users whose k^{out} is larger than a specified value) are shown in Figure 4(a) and 4(b) for the type 1 and 2 user, respectively. Figure 4 indicates that a follower of a type 2 user tends to have larger k^{out} than a follower of a type 1 user on average. For the Japanese, Russian, and Korean groups, the mean $\pm$ standard deviation is equal to 1125 ± 7193 for type 1 and 20070 ± 48849 for type 2, 1526 ± 11435 for type 1 and 9068 ± 31711 for type 2, 4119 ± 11316 for type 1 and 20114 ± 40424 for type 2, respectively.

Because k^{out} obeys relatively long-tailed distributions (Figure 4), the comparison of the mean values is insufficient. Therefore, we quantify the classification performance of the follower's k^{out} by using the receiver operating characteristic curve (ROC) based on the two distributions of k^{out} for each language (Tuffery 2011). The ROC is the trajectory of the false positive (i.e., fraction of type 2 users that are mistakenly judged as type 1 on the basis of k^{out}) and the true positive (i.e., fraction of type 1 users correctly judged as type 1 with the same threshold), when the threshold for classification is varied. The area under the curve (AUC) of the ROC falls between 0.5 and 1. When AUC is large, the two distributions are well separated such that users are accurately judged as type 1 or 2. The values of AUC for different

language groups are shown in Table 4. The AUC is larger for the Japanese, Russian, and Korean groups than for the other four groups. It should be noted that for the Japanese, Russian, and Korean groups, the type 1 and type 2 users are more clearly distinguished than for the other groups (Table 2).



(a)



(b)

Figure 4: Survival function of the number of friends (i.e., k^{out}) for the followers of a (a) type 1 user and (b) type 2 user.

Table 4: AUC values for the follower's k^{out} and the follower' reciprocity.

Language	Follower's k^{out}	Follower's reciprocity
English	0.680	0.816
Spanish	0.705	0.740
Japanese	0.831	0.838
Portuguese	0.628	0.681
Russian	0.819	0.875
Korean	0.748	0.796
French	0.722	0.883

Follower's reciprocity Third, we measure the number of reciprocal links owned by a follower of a type 1 or 2 user, divided by k^{out} for this follower (Figure 3(b)). We call the ratio the follower's reciprocity, which ranges between 0 and 1. If the follower's reciprocity is large, the follow that a type 1 or 2 user receives may not be valuable in the sense that the follower easily establishes reciprocal links with others, perhaps to advertise themselves (Ghosh et al. 2012) or mutually connect with close friends.

To calculate the follower's reciprocity and also the fourth quantity C_i described below, we have to acquire IDs of the followers and friends for each user following a type 1 or 2 user. This operation requires too much time because we can call API resources a limited number of times per hour. Therefore, we calculate the quantity of interest (follower's reciprocity or C_i) for randomly selected 100 users following each type 1 or 2 user.

We found that followers of type 2 users have larger follower's reciprocity values than followers of type 1 users on average. This holds true in particular for the Japanese (0.434 $\pm$ 0.250 for type 1 versus 0.762 $\pm$ 0.224 for type 2, where the mean and standard deviation are calculated on the basis of all the users that follow any of the ten randomly selected type 1 or 2 users), Russian (0.231 $\pm$ 0.287 for type 1 versus 0.703 $\pm$ 0.266 for type 2), and Korean (0.491 $\pm$ 0.352 for type 1 versus 0.846 $\pm$ 0.206 for type 2) groups. Because the follower's reciprocity in fact obeys a rather long tailed distribution, as in the case of the second quantity, we calculate the AUC for the follower's reciprocity. The AUC values for the seven language groups are shown in Table 4. The AUC is relatively large such that the follower's reciprocity is effective at distinguishing between type 1 and 2 users.

Local clustering coefficient Fourth, we examine the local clustering coefficient (Newman 2010), denoted by C_i for type 1 or 2 user labeled i, which is the density of triangles including user i. Because the Twitter followership network has a large global clustering coefficient (Romero

and Kleinberg 2010), a considerable portion of users would have large C_i, and C_i is expected to serve to characterize users. For a type 1 or 2 user i having indegree $k_i^{\rm in}$, there can be maximum $k_i^{\rm in}(k_i^{\rm in}-1)/2$ triangles that include user i, whereby we impose that two followers of i are connected by reciprocal links to be qualified as a triangle including i. We define C_i as the actual number of triangles divided by $k_i^{\rm in}(k_i^{\rm in}-1)/2$. By definition, C_i ranges between 0 and 1. If C_i is large, the follow that a type 1 or 2 user i receives may be not as valuable as otherwise because the user is likely to be followed by many similar users, where the similarity is implicit in reciprocal links between the followers.

As shown in Table 5, C_i is significantly larger for type 2 users than type 1 users except for the Portuguese group. It should be noted that the difference is prominent for the Japanese, Russian, and Korean groups, for which the distinction between the type 1 and type 2 users are clear.

Table 5: Local clustering coefficient. The mean and standard deviation are calculated on the basis of ten randomly selected users of each type and language.

Language	Type 1	Type 2
English	0.0036±0.0087	0.0293±0.0275
Spanish	0.0017±0.0016	0.0098±0.0077
Japanese	0.0039±0.0039	0.1334±0.0875
Portuguese	0.0025±0.0034	0.0214±0.0417
Russian	0.0086±0.0110	0.0919±0.0359
Korean	0.0988±0.1505	0.3648±0.2197
French	0.0021±0.0027	0.0419±0.0341

Abundance of type 2-like users among followers Fifth, we define the fraction of type 2-like users among the followers. It should be noted that $k^{\rm out}$ of the followers (second quantity that we have investigated) and the follower's reciprocity (third quantity) also capture the tendency that users following a type 1 or 2 user resemble type 2 users to some extent. Here we define a more direct measure called the fraction of type $2'$ users as the fraction of followers of a type 1 or 2 user satisfying $k^{\rm out}/1.1 \le k^{\rm in} \le 1.1 \times k^{\rm out}$. Similar to the definition of d, we exclude the followers with $k^{\rm in}$ and $k^{\rm out}$ values smaller than a prescribed threshold from the calculation of the fraction of type $2'$ users. The analysis of the four quantities carried out above suggests that the follow that a type 2 user receives is probably less valuable than that a type 1 user receives. If we accept this assumption, a large fraction of type $2'$ users among the followers of type 2 users as compared to among the followers of type 1 users would lend another support to our claim that the follow that a type 2 user receives is not as valuable as that a type 1 user receives. For each user type and language, we calculate the mean and standard deviation of the fraction of type $2'$ users on the basis of the ten randomly selected users.

The results with the threshold equal to 100 (i.e., followers having $k^{\rm in}$, $k^{\rm out} \le 100$ are excluded from the calculation of the fraction of type $2'$ users) and 2000 are shown in Table 6. The table indicates that type 2 users are significantly more likely to be followed by type $2'$ users than type 1 users are. This tendency is stronger for the Japanese, Russian, and Korean groups than the other four language groups.

PageRank of the two types of users

We estimate their PageRank because all the quantities measured in the previous sections are local ones, whereas the PageRank quantifies global importance of a node. To quantify importance of nodes in directed networks, the PageRank algorithm is often used (Brin and Page 1998; Langville and Meyer 2006). In fact, the PageRank and its variants have also been used for ranking users in Twitter social networks (Weng et al. 2010). By definition, the PageRank of a user would be small if the user's follower has a large $k^{\rm out}$(i.e., number of friends). Therefore, we expect that a type 1 user in general has a larger PageRank value than a type 2 with the same number of followers. The PageRank of a node is proportional to the frequency with which a random walker visits the node. The walker is defined to move to one of downstream neighbors with the equal probability $(1-q)/k^{\rm out}$ such that the total probability of such an ordinary random walk is equal to $1-q$. With the remaining probability q, the walker jumps to an arbitrary node with the equal probability, which is the so-called teleportation. Although the PageRank is often strongly correlated with $k^{\rm in}$ (Fortunato et al. 2008; Ghoshal and Barabási 2011), it is not always the case (Donato et al. 2004; Masuda and Ohtsuki 2009). For Twitter networks, it was reported that $k^{\rm in}$ (i.e., number of followers) and the PageRank are strongly correlated (Kwak et al. 2010).

In the following, we compare the PageRank of type 1 and type 2 users. Because the exact calculation of the PageRank requires the full information about the connectivity of the network, we approximate the PageRank by emulating the random walk. We first select a user with the equal probability from the set of users. The random walk starts from the selected user. We selected the initial position of the random walk from the set of Japanese users collected by the random sapling. We confined ourselves to Japanese users because the distinction between type 1 and 2 users is clear for them. Second, we move to a friend of the selected user with the equal probability $1/k^{\rm out}$. Third, we repeat the same random hopping ten times. If the walker hits a user without any follower, we terminate the random walk. Finally, we redraw a starting user without replacement and carry out the ten-step random walk for 1500 randomly selected initial nodes. Stopping the random walk after ten steps corresponds to the teleportation with probability $q = 1/11$. This value is comparable with the conventional teleportation probability $q = 0.15$ (Brin and Page 1998; Langville and Meyer 2006). The probability that the walker hits a given type 1 or 2 user is very small. To enhance the probability that the walker hits any of type 1 or 2 users, we increased the number of type 1 users and that of type 2 users as follows. First, we focused on type 1 and 2 Japanese users identified by the neighbor sampling because it is much rarer to find a type 1 or 2 user with the random sampling. Second, we added two Japanese seed users. We carried out the neighbor sampling with the two seed users to find new type 1 and 2 users employed as additional targets of the random

Table 6: Fraction of type $2'$ users for different user types and languages.

Language	Type 1 (threshold=100)	Type 2 (threshold=100)	Type 1 (threshold=2000)	Type 2 (threshold=2000)
English	0.055±0.046	0.244±0.112	0.212±0.115	0.402±0.099
Spanish	0.022±0.008	0.123±0.045	0.057±0.059	0.357±0.078
Japanese	0.122±0.049	0.486±0.141	0.326±0.121	0.674±0.065
Portuguese	0.022±0.012	0.108±0.137	0.071±0.043	0.213±0.149
Russian	0.091±0.089	0.397±0.055	0.279±0.174	0.603±0.045
Korean	0.313±0.353	0.758±0.192	0.506±0.313	0.912±0.047
French	0.034±0.025	0.248±0.109	0.132±0.065	0.449±0.136

walk.

Because the PageRank is usually correlated with k^{in}, we counted the number of visits to type 1 or 2 users for each of the four groups defined by different k^{in} ranges (Table 7). For each degree group, the walker visits type 1 users more frequently than type 2 users. Type 1 users are more important than type 2 users in terms of the PageRank.

Table 7: Frequency that the random walker visits type 1 or 2 users. For each degree group defined by a distinct range of k^{in}, we found less type 1 users than type 2 users by the neighbor sampling. Therefore, we randomly sampled users from the set of type 2 users such that the number of type 2 users is equal to that of type 1 users (e.g., 941).

k^{in}	Number of users	Type 1	Type 2
2500–7500	941	43	12
7500–12500	224	16	4
12500–17500	93	10	4
17500–22500	62	10	3

Discussion and Conclusions

By measuring different network-based quantities, we showed that type 1 and 2 users have different network properties although they have comparably many followers. On average, type 1 users, defined by a small number of friends, are characterized by less reciprocal links, possession of followers with less reciprocal links and less friends, and larger PageRank values, than type 2 users. The distinction between the two types is stronger in the Japanese, Russian, and Korean language groups than the English, Spanish, Portuguese, and French groups. Warning that a specific user is of type 1 or 2 may help promote social etiquette on Twitter.

Some of the type 1 and 2 users that we sampled were admittedly spammers, organizational accounts, and bots. Nevertheless, according to our visual inspection, there were few of them, in particular among the Japanese and Spanish groups. Because of their small fractions, we believe that the effects of the spammer-type accounts on our results are limited.

User IDs suspected of organized link farming activities may follow other users and anticipate that they are followed back. Such users may be the so-called social capitalists, who aim to promote their legitimate contents to be broadcast to wide audience (Ghosh et al. 2012). They tend to exchange reciprocal links with others and are densely connected with each other. Similar to social capitalists, spam followers also tend to have high reciprocity. These behavioral properties of social capitalists are consistent with the high reciprocity and homophily of type 2 users found in the present study. However, analysis of the intention and behavior of the type 2 users is beyond the scope of the present study; we analyzed the followership networks but not the contents or propagation of tweets. It should be also noted that, unlike Ghosh et al. (Ghosh et al. 2012), we did not look at connectivity of users to spams. Type 2 users may exchange links as a part of link farming activities, spam activities, or just to assure mutual friendship.

Ghosh et al. cite celebrities and popular bloggers as examples of social capitalists (Ghosh et al. 2012). However, our manual inspection of the users' profiles suggests that more celebrities and popular bloggers are found among type 1 rather than type 2 users. They also conclude that social capitalists and spammers are influencers (Ghosh et al. 2012). In contrast, our type 2 users would have much smaller influences in terms of the PageRank than type 1 users. Although the reason for this discrepancy is unclear, our main claim is that we can classify seemingly influential (i.e., having large number of followers) users into rather discrete two types. Social capitalists identified in (Ghosh et al. 2012) may be a mixture of type 1 and 2 users. To subcategorize the social capitalists into type 1 and type 2 -like classes by incorporating the information about tweets and connectivity to spams is warranted for future work.

The number of followers and that of friends were very close for most users in a previous report (Weng et al. 2010). The results are inconsistent with ours; we found that the proximity depends on users (Figure 1) and the language (Table 2). The reason why type 1 users were not found in the previous study (Weng et al. 2010) is unclear but may be that they mainly investigated English speaking users.

Weng et al. proposed the TwitterRank to rank users (Weng et al. 2010). The TwitterRank is different from the PageRank because in the former the walker tends to transit to a friend that is similar to the user and tweets many times on each topic. The TunkRank is another variant of the PageRank in which the retweet probability is taken into account in determining the transition probability (*http://tunkrank.com/*). In

the present work, we used the original PageRank without taking these non-network features into account. Our aim was to extract the information about the value of users only on the basis of the network structure. Better characterizing different types of users by combining the present method with users' activities is an obvious future question.

Web Ecology project measures the influence of the user on the basis of the activities received by the user, which include the number of retweets divided by that of tweets (Leavitt et al. 2010). Our results are in line with this definition because a network equivalent of their measure is given by $k^{\mathrm{in}}/k^{\mathrm{out}}$, which is much larger than unity for type 1 users and approximately equal to unity type 2 users.

Acknowledgment

We thank Katsumi Sakata for discussing ideas leading to this project, and Mitsuhiro Nakamura, Ryosuke Nishi, and Taro Takaguchi for critical reading of the manuscript. We acknowledge the financial support provided through Grants-in-Aid for Scientific Research (No. 23681033) from MEXT, Japan, and JST ERATO Kawarabayashi Large Graph Project.

References

[Bakshy et al. 2011] Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 65–74.

[Brin and Page 1998] Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30:107–117.

[Cha et al. 2010] Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media*, 10–17.

[Donato et al. 2004] Donato, D.; Laura, L.; Leonardi, S.; and Millozzi, S. 2004. Large scale properties of the webgraph. *European Physical Journal B* 38:239–243.

[Fortunato et al. 2008] Fortunato, S.; Boguñá, M.; Flammini, A.; and Menczer, F. 2008. Approximating PageRank from in-degree. *Lecture Notes in Computer Science* 4936:59–71.

[Ghosh et al. 2012] Ghosh, S.; Viswanath, B.; Kooti, F.; Sharma, N. K.; Korlam, G.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. P. 2012. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, 61–70.

[Ghoshal and Barabási 2011] Ghoshal, G., and Barabási, A. L. 2011. Ranking stability and super-stable nodes in complex networks. *Nature Communications* 2:394.

[Java et al. 2007] Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 56–65.

[Kwak et al. 2010] Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. ACM.

[Langville and Meyer 2006] Langville, A. N., and Meyer, C. D. 2006. *Google's PageRank and beyond: The Science of Search Engine Rankings*. Princeton University Press.

[Leavitt et al. 2010] Leavitt, A.; Burchard, E.; Fisher, D.; and Gilbert, S. 2010. The influentials: new approaches for analyzing influence on twitter. In Web Ecology Project, http://www.webecologyproject.org/.

[Masuda and Ohtsuki 2009] Masuda, N., and Ohtsuki, H. 2009. Evolutionary dynamics and fixation probabilities in directed networks. *New Journal of Physics* 11:033012.

[Newman 2010] Newman, M. E. J. 2010. *Networks: An introduction*. Oxford University Press, Inc.

[Romero and Kleinberg 2010] Romero, D. M., and Kleinberg, J. 2010. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 138–145.

[Russell 2011] Russell, M. 2011. *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. O'Reilly Media.

[Suh et al. 2010] Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 177–184.

[Takhteyev, Gruzd, and Wellman 2012] Takhteyev, Y.; Gruzd, A.; and Wellman, B. 2012. Geography of twitter networks. *Social Networks* 34:73–81.

[Tuffery 2011] Tuffery, S. 2011. *Data Mining and Statistics for Decision Making*. Wiley.

[Weng et al. 2010] Weng, J.; Lim, E. P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 261–270.